4. *Менщиков А.А., Перфильев В.Э., Федосенко М.Ю., Фабзиев И.Р.* Основные проблемы использования больших данных в современных информационных системах // Столыпинский вестник. – 2022. – Т. 4. № 1. – URL: https://stolypin-vestnik.ru/wp-content/uploads/2022/01/31.pdf (дата обращения 16.09.2024).

5. *Evsutin O., Ivanov F., Dzhanashia K.* Watermarking for social networks images with improved robustness through polar codes // IEEE Access. – 2024. – Vol. 12. – P. 118154-118168.

6. *Заславская В.Л.* Проблемы использования слабоструктурированных и неструктурированных данных в системах бизнес-аналитики // Мягкие измерения и вычисления. – 2022. – Т. 59. № 10. – С. 84-94.

Домашкин А.Д., Логинова Л.Н.

Сравнительный анализ алгоритмов машинного обучения для обнаружения аномалий в информационных системах

Аннотация: В работе рассматривается сравнительный анализ алгоритмов машинного обучения для обнаружения аномалий в информационных системах, включая такие методы, как K-ближайших соседей, метод опорных векторов, кластеризация K-средних и изолированные деревья. Подчеркивается важность выбора подходящего алгоритма в зависимости от типа данных и требований к точности для повышения качества информационной безопасности.

Ключевые слова: информационные системы, машинное обучение, аномалии, информационная безопасность, задачи классификации

В условиях стремительного увеличения объема данных и сложности информационных систем (ИС) обеспечение их безопасности становится ключевым вопросом. Число кибератак только на российские ИТ-компании, которые используют современные методы защиты, во втором полугодии 2023 года выросло в 4 раза по сравнению с аналогичным периодом 2022-го и

достигло 4 тыс. [1]. Это означает, что современные системы обнаружения аномалий (СОВ) могут пропустить аномалии и подозрительный трафик. Для совершенствования СОВ необходимо использовать эффективные методики и технологии. Одной из таких технологий является машинное обучение (МО) [2].

Важным шагом на пути к повышению качества обеспечения информационной безопасности (ИБ) является обнаружение аномалий – непредвиденных отклонений от нормы, которые могут указывать на атаки, сбои в работе систем или другие инциденты. Известны алгоритмы машинного обучения, которые способны автоматически анализировать большие объемы данных и выявлять скрытые аномалии. Так, для обнаружения отклонений поведения ИС разработаны различные методики, однако выбор эффективного алгоритма зависит от многих факторов, таких как тип данных, требуемая точность и вычислительные ресурсы.

ИС могут сталкиваться с аномалиями разных видов: от нетипичных действий пользователей до намеренных попыток злоумышленников скомпрометировать конфиденциальные данные. В последнее время всё чаще используются технологии МО, которые помогают не только обнаружить уже известные виды атак, но и выявлять новые, не изученные угрозы.

Алгоритмы обнаружения аномалий можно разделить на несколько категорий [3]:

– обучение с учителем (Supervised Learning): применяется в тех случаях, когда доступны метки нормальных и аномальных данных;

– обучение без учителя (Unsupervised Learning): используется при отсутствии меток, при этом требуется алгоритмическое выявление аномалий на основе анализа структуры данных;

– полуобучение (Semi-supervised Learning): предполагает наличие небольшого количества данных с метками, на основе которых строится модель, способная выявлять аномалии в новых данных.

Каждая из перечисленных выше категорий включает в себя алгоритмы МО, которые могут быть использованы для обнаружения отклонений в ИС. К алгоритмам МО относятся [3]:

1) K-ближайших соседей (K-Nearest Neighbors, KNN) [3],
2) метод опорных векторов (Support Vector Machines, SVM),

3) алгоритм кластеризации K-средних (K-Means),

4) линейный дискриминантный анализ (Linear Discriminant Analysis, LDA),

5) алгоритм леса изолированных деревьев (Isolation Forest).

Рассмотрим более подробно каждый из них.

1. Алгоритм KNN – это один из наиболее простых методов МО, который базируется на идее классификации новых объектов на основе их близости к уже известным объектам. В контексте обнаружения отклонений данный метод позволяет классифицировать точку как аномальную, если она находится вдали от большинства других точек, которые считаются нормальными.

Преимущества KNN:

– прост в реализации и понятен в интерпретации результатов,

– эффективен для небольших наборов данных.

Недостатки KNN:

– неэффективно масштабируется для больших объемов данных,

– требует выбора оптимального числа ближайших соседей (K),

– чувствителен к выбору метрик расстояния.

Данный метод активно применяется в задачах по детектированию аномального сетевого трафика и его классификации. К решаемым задачам с помощью алгоритма KNN относятся задачи выявления атак, классификации трафика, выявления спама, фишинга и вредоносных программ [4].

2. Метод опорных векторов успешно применяется для классификации и обнаружения аномалий. Для их поиска используется модификация метода SVM, называемая One-Class SVM, которая обучается только на нормальных данных и отделяет их от аномальных [3].

Преимущества SVM:

– высокая точность при правильной настройке параметров,

– эффективен для высокоразмерных данных.

Недостатки SVM:

– требует тщательной настройки гиперпараметров (например, параметра ядра),

– затруднена интерпретация модели,

– вычислительные затраты при работе с большими объемами данных.

Примером применения алгоритма SVM является выявление аномалий в логах ИС [5].

3. Алгоритм K-Means используется для разделения данных на кластеры, при этом аномалии могут быть определены как точки, находящиеся вдали от центроидов кластеров. Метод стал популярным инструментом для выявления отклонений в случае, когда данные имеют естественные группы [3].

Преимущества K-Means:

– прост в реализации,
– эффективен при больших объемах данных.

Недостатки K-Means:

– требует заранее определить количество кластеров,
– чувствителен к шуму и выбросам,
– модель может быть неустойчива к вариациям в данных.

4. LDA – это алгоритм, который использует линейные комбинации признаков для разделения классов данных [3]. В случае обнаружения отклонений он может быть применен для выделения нормальных данных и аномалий на основе геометрических характеристик данных.

Преимущества LDA:

– прост и эффективен на линейно разделимых данных,
– легко интерпретируется.

Недостатки LDA:

– неэффективен на нелинейных данных,
– ограниченно применим для высокоразмерных данных с небольшим числом наблюдений.

5. Isolation Forest – один из специализированных методов для обнаружения аномалий [3]. В отличие от других алгоритмов он не пытается моделировать нормальные данные, а изолирует отклонения на основе их уникальных характеристик.

Преимущества Isolation Forest:

– эффективен при работе с большими наборами данных,
– не требует предположений о распределении данных.

Недостатки Isolation Forest:

– может дать ложные результаты на сильно искаженных наборах данных,
– неэффективен на небольших выборках.

Алгоритм Isolation Forest применяется для детектирования аномалий в поведении пользователей ИС, в поведении операционной системы, в системах мониторинга, в сетевом трафике.

Для проведения сравнительного анализа вышеперечисленных алгоритмов МО были использованы следующие критерии.

1. Точность обнаружения аномалий – способность алгоритма корректно идентифицировать отклонения.

2. Масштабируемость – эффективность работы метода с большими объемами данных.

3. Скорость работы – время, затрачиваемое на обучение и прогнозирование.

4. Интерпретируемость результатов – возможность объяснения разделения данных на нормальные и аномальные.

5. Чувствительность к гиперпараметрам – зависимость результата от правильной настройки гиперпараметров.

На основе приведенных критериев был произведен сравнительный анализ алгоритмов МО, при этом использовались одинаковые выборки данных для алгоритмов. Выборки данных были собраны с автоматизированных рабочих мест сотрудников и включали в себя: сетевую активность устройства, активность операционной системы, системные вызовы и запуск процессов. Шкала оценки была сформирована из среднего значения, полученного по результатам выполнения работы соответствующего алгоритма поиска аномалий, где «Низкая» – это значение от 0 до 2, «Средняя» – от 3 до 4, «Высокая» – 5.

В таблицах 1 и 2 представлены результаты исследования.

Таблица 1 – Анализ алгоритмов (начало)

Алгоритм	**Точность**	**Масштабируемость**	**Скорость**
K-ближайших соседей (KNN)	Средняя	Низкая	Низкая
One-Class SVM	Высокая	Низкая	Средняя
K-средних (K-Means)	Средняя	Высокая	Высокая
Линейный дискриминантный анализ (LDA)	Средняя	Средняя	Высокая
Isolation Forest	Высокая	Высокая	Высокая

Таблица 2 – Анализ алгоритмов (продолжение)

Алгоритм	**Интерпретируемость**	**Чувствительность к гиперпараметрам**
K-ближайших соседей (KNN)	Высокая	Высокая
One-Class SVM	Низкая	Высокая
K-средних (K-Means)	Средняя	Средняя
Линейный дискриминантный анализ (LDA)	Высокая	Средняя
Isolation Forest	Средняя	Низкая

Результаты исследования показали, что использование алгоритма Isolation Forest для выявления аномалий является наиболее эффективным среди остальных алгоритмов.

Авторами были рассмотрены несколько популярных алгоритмов МО, каждый из которых имеет свои преимущества и недостатки. Выбор эффективного алгоритма зависит от характеристик данных и конкретных условий применения. Дальнейшие исследования направлены на разработку более эффективных и устойчивых методов обнаружения отклонений, которые будут учитывать динамическую природу угроз в современных ИС.

Литература:

1. Число кибератак в России и в мире. – URL: https://www.tadviser.ru/index.php/Статья:Число_кибератак_в_России_и_в_мире?ysclid=m2p4wkbebz321154872 (дата обращения 10.09.2024).

2. *Хромов С.К., Кулагин М.А., Сидоренко В.Г.* Автоматизация сопровождения пользователей автоматизированных систем управления на базе машинного обучения / Новые информационные технологии в исследовании сложных структур: материалы Тринадцатой Международной конференции, Томск, 07-09 сентября 2020 года. – Томск: Национальный исследовательский Томский государственный университет, 2020. – С. 74-75.

3. The Cylance Data Science Team. Introduction to artificial intelligence for security professionals. – Irvine, CA: The Cylance Press, 2017. – 155 p.

4. *Кусакина Н.М.* Построение CNN для решения задачи выявления аномалий сетевого трафика // Перспективы науки. – 2019. – № 6(117). – С. 53-55.

5. *Кирячек В.А.* Разработка ML-подхода идентификации аномалий по логам компьютерных систем с помощью методов обработки естественного языка / Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: Материалы Всероссийской конференции с международным участием, Москва, 08-12 апреля 2024 года. – М.: Российский университет дружбы народов им. П. Лумумбы, 2024. – С. 485-490.